International Journal of Enhanced Research in Science Technology & Engineering, ISSN: 2319-7463
Vol. 4 Issue 6, June-2015, pp: (106-110), Impact Factor: 1.252, Available online at: www.erpublications.com

Supervised Multilayer Perceptron Network based Decision Tree Algorithm for Cancer Classification

Gursharan Singh[1], Dr. Kulwinder Singh Mann[2]
Research Scholar[1], Associate Professor[2]
[1,2]Department of Information Technology, Guru Nanak Dev Engineering College, India

Abstract: Decision tree based techniques are being continuously evolved for classification of medical datasets .Different algorithms based on soft computing and hard computing have been developed to apply on medical datasets. In current research, an intelligent technique inspired from neural network, supervised multilayer perceptron decision tree have been applied and evaluated successfully to classify lung cancer based datasets. A comparison has been made with other techniques to check the effectiveness of the proposed method. Simulation results shows that proposed technique achieved 100% accuracy to classify cancer data sets which is more as compared other techniques.TP rate, ROC and Precision is highest for proposed method amongst other method. Hence, proposed algorithm is optimum to classify cancer related datasets.

Keywords: MLPT, NBT, SVMT, SMLPT.

1. Introduction

Data mining[1] is the process of digging data for discovering latent patterns which can be translated into valuable information. Data mining usage witnessed unprecedented growth in the last few years. Of late the usefulness of data mining techniques has been realized in Healthcare domain. This realization is in the wake of explosion of complex medical data. Medical data mining can exploit the hidden patterns present in voluminous medical data which otherwise is left undiscovered. Data mining techniques which are applied to medical data include association rule mining for finding frequent patterns, prediction, classification and clustering. Traditionally data mining techniques were used in various domains. However, it is introduced relatively late into the Healthcare domain. Nevertheless, as on today lot of research is found in the literature. This has led to the development of intelligent systems and decision support systems in Healthcare domain for accurate diagnosis of diseases, predicting the severity of various diseases, and remote health monitoring.[18] Especially the data mining techniques are more useful in predicting heart diseases, lung cancer, and breast cancer and so on. The data mining techniques that have been applied to medical data include Apriori and FP Growth, [19] unsupervised neural networks, linear genetic programming, Association rule mining, Bayesian Ying Yang , decision tree algorithms like ID3, C4.5, C5, and CART , outlier prediction technique , [20]Fuzzy cluster analysis, classification algorithm, Bayesian Network algorithm, Naive Bayesian, combination of K-means, Self Organizing Map (SOM) and Naïve Bayes, [21]Time series technique, combination of SVM, ANN and ID3, clustering and classification, SVM, , FCM, k-NN, and Bayesian Network.

II. Literature Survey

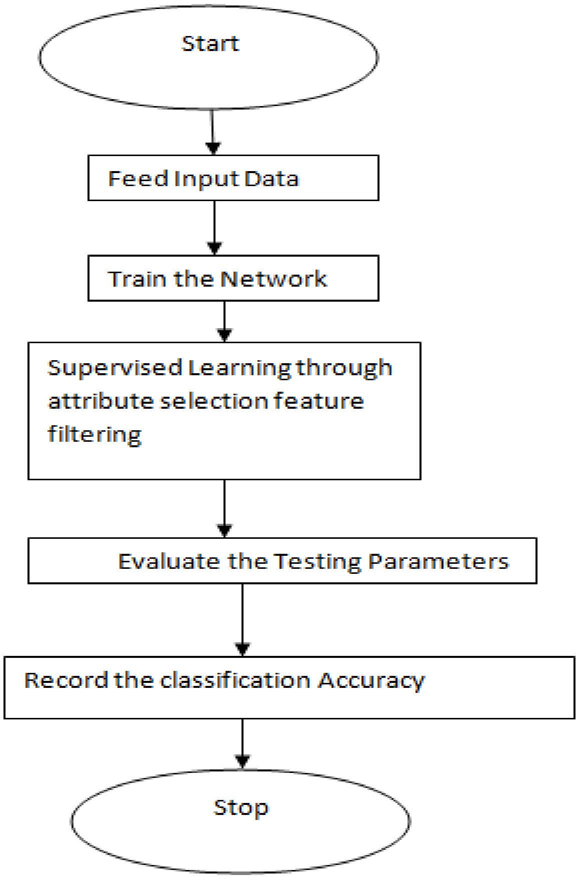
Umar et al[1] applied data mining techniques for birth outcomes. **Cong et al. [2]** stated that hereditary syndromes can be detected automatically using data mining techniques. **Hai et al. [3]** discussed medical data mining through unsupervised neural networks besides a method for data visualization. They also emphasized the need for preprocessing prior to medical data mining. In the year 2000 **Carshen et al [4]** , bioengineering professor, identified the need for data mining methods to mine medical multimedia content. **Shariq[5]** identified problems in medical data mining. The problems include missing values, data storage with respect to temporal data and multi-valued data, different medical coding systems being used in Hospital Information Systems (HIS). **Sunil [6]** explored and analyzed two programming models such as neural networks, and linier genetic programming for medical data mining. **Thanh et al [7]** proposed and implemented a symbolic rule extraction workbench for generating emerging rule-sets. **Xiang et al. [8]** explored the usage of rule-sets as results of data mining for building rule-based expert systems. Markus et al [9] proposed an algorithm for extracting association rules from medical image data. The association rule mining discovers frequently occurring items in the given dataset. **Doron et al** [10] proposed a classification method based on Bayesian Ying Yang (BYY) which is a three layered model. They applied this model to classify liver disease through automatic discovery of medical trends. **Adepele et al. [11]** proposed architecture for mining geno-medical data in heterogeneous and grid-based distributed infrastructures. **Cindy et al. [12]** focused on decision tree data mining algorithm for medical image analysis. Especially they studied on lung cancer diagnosis through classification of x-ray images. **Jeong et al. [13]** presented an outlier prediction method for improving performance of classification as part of medical data mining. **Jann et al.[14]** applied fuzzy cluster analysis for medical images. They used decision tree algorithm to classify mammography into normal and abnormal cases. **Safwan et al. [15]** applied classification algorithm to diagnose cardio vascular diseases. For classification effectiveness they focused on two feature extraction techniques namely automatic feature selection and expert judgment. **Yanwai et al. [16]** introduced web based data mining for the application of telemedicine. Tsang **et al. [17]** presented an approach to integrate PSO rule mining methods and classifier on patient dataset. They used Particle Swarm Optimization technique as well. The results revealed that, their approach is capable of performing surgery candidate selection process effectively in epilepsy.

III. Proposed Method

- **Supervised Multilayer Perceptron Tree**

A supervised multilayer perceptron tree (SMLPT) is a trained feed forward neural network model that maps sets of input data onto a set of appropriate outputs. A SMLPT consists of multiple layers of nodes in a directed graph, with each layer fully connected to the next one. Except for the input nodes, each node is a neuron (or processing element) with a nonlinear activation function. SMLPT utilizes a supervised learning technique called back propagation for training the network through attribute selection feature. SMLPT is a modification of the standard multilayer perceptron and can distinguish data that are not linearly separable.

- **Flow Chart**


Start
Feed Input Data
Train the Network
Supervised Learning through attribute selection feature filtering
Evaluate the Testing Parameters
Record the classification Accuracy
Stop

IV. Cancer Data Set

Relation: lung-cancer
Instances: 32
Attributes: 57

V. Computation Time for different Decision Tree Algorithms

1) Naïve Bayes Tree
Time taken to build model: 0.02 seconds

2) Multilayer Perceptron Tree
Time taken to build model: 8.16 seconds

3) Support Vector Machine Tree
Time taken to build model: 0.48 seconds

4) Supervised MultiLayer Perceptron Tree
Time taken to build model: 7.83 seconds

VI. Simulation Results

Comparative analysis is done for checking the effectiveness of the proposed method. As observed in Table 1., we can see that proposed method SMLPT is having high classification rate with accuracy of 100% as compared to other decision tree algorithms and less error rate. It achieves high value of testing parameters (TP-True Positive, FP-False Positive, ROC-Region of Curve).

Table 1. Comparative Analysis of Error Parameters and Accuracy

Technique	Kappa statistic	Mean absolute error	Root mean squared error	Relative absolute error	Root relative squared error	Accuracy %
Naïve Bayes Decision Tree	0.44	0.2376	0.4702	57.5927	104.0271	78.125
Multi Layer Perceptron	0.12	0.3247	0.5302	78.7205	117.2976	65.625
Support Vector Machine	0.12	0.3438	0.5863	83.3333	129.7132	65.625
Supervised MLP	**1**	**0.0038**	**0.0047**	**92.26**	**1.0416**	**100**

Table 2. Accuracy by Class with Naïve Bayes Tree (NBT)

Class	Recall	TP Rate	FP Rate	Precision	F-Measure	ROC Area
Non-Cancerous	0.556	0.556	0.13	0.625	0.588	0.773
Cancerous	0.87	0.87	0.444	0.833	0.851	0.773
Weighted Avg.	0.781	0.781	0.356	0.775	0.777	0.773

Table 3. Accuracy by Class with Multilayer Perceptron Tree (MLPT)

Class	Recall	TP Rate	FP Rate	Precision	F-Measure	ROC Area
Non-Cancerous	0.851	0.851	0.624	0.763	0.805	0.691
Cancerous	0.376	0.376	0.149	0.516	0.435	0.691
Weighted Avg.	0.71	0.71	0.483	0.69	0.695	0.691

Table 4. Accuracy by Class with Support Vector Machine Tree (SVMT)

Class	Recall	TP Rate	FP Rate	Precision	F-Measure	ROC Area
Non-Cancerous	0.866	0.866	0.706	0.744	0.8	0.637
Cancerous	0.294	0.134	0.149	0.481	0.365	0.637
Weighted Avg.	0.696	0.536	0.483	0.665	0.671	0.637

Table 5. Accuracy by Class with Supervised Multilayer Perceptron Tree (SMLPT)

Class	Recall	TP Rate	FP Rate	Precision	F-Measure	ROC Area
Non-Cancerous	1	1	0	1	1	1
Cancerous	1	1	0	1	1	1
Weighted Avg.	1	1	0	1	1	1

From Table 5, we can see that True positive rate ,precision of solution, F measure and ROC value is highest for proposed technique SMLPT as compared to other methods as shown in table 2,3,4. Hence, proposed method SMLPT is effective in testing parameters.

Conclusion

The proposed SMLPT based approach evolved as optimal approach to classify the cancer datasets with a remarkable accuracy of 100% and fast computation time of 7.83 seconds as compared to MLP technique and other classification methods. With such high accuracy in proposed method, it will be easy to identify the cancer and non cancer patients from different attributes of people for large data chunks where other decision tree algorithms fail to achieve high accuracy.

Acknowledgement

The authors would like to thank the management of GNDEC for providing facilities to carry out this work.

References

[1]. Umair Abdullah (2008). "Analysis of Effectiveness of Apriori Algorithm in Medical Billing Data Mining". Proceedings of IEEE, pp.1-5.

[2]. Cong-Rui Ji and Zhi-Hong Deng. (2009). Mining Frequent Ordered Patterns without Candidate Generation. Proceedings of IEEE, pp.1-5.

[3]. Hai-Tao He and Shi-Ling Zhang. (2007). "A New method for Incremental Updating Frequent patterns mining", Proceedings of IEEE, pp.1-4.

[4]. Carson Kai-Sang Leung, Christopher L. Carmichael and Boyu Hao. (2007). "Efficient Mining of Frequent Patterns from Uncertain Data",. Proceedings of IEEE ,pp.489-494.

[5]. Shariq Bashir, Zahid Halim, A. Rauf Baig. (2008).," Mining Fault Tolerant Frequent Patterns using Pattern Growth Approach". Proceedings of IEEE ,pp.172-179.

[6]. Sunil Joshi and Dr. R. C. Jain. (2010). "A Dynamic Approach for Frequent Pattern Mining Using Transposition of Database", Proceedings of IEEE, pp.498-501.

[7]. Thanh-Trung Nguyen. (2010)."An Improved Algorithm for Frequent Patterns Mining Problem", Proceedings of IEEE, pp.503-507.

[8]. Xiaoyong Lin and Qunxiong Zhu. (2010). "Share-Inherit: A novel approach for mining frequent patterns", Proceedings of IEEE, pp.2712-2717.

[9]. Markus Brameier and Wolfgang Banzhaf. (2001)."A Comparison of Linear Genetic Programming and Neural Networks in Medical Data Mining", Proceedings of IEEE ,pp.1-10.

[10]. Doron Shalvi and Nicholas DeClaris., (2008). "An Unsupervised Neural Network Approach to Medical Data Mining Techniques.", Proceedings of IEEE , pp.1-6.

[11]. Adepele Olukunle and Sylvanus Ehikioya, (2009). "A Fast Algorithm for Mining Association Rules in Medical Image Data", Proceedings of IEEE,. pp.1-7.

[12]. Cindy L. Bethel and Lawrence O. Hall and Dmitry Goldgof (2007). "Mining for Implications in Medical Data." , Proceedings of IEEE , pp.1-4.

[13]. Jeong-Yon Shim, Lei Xu (2009). "Medical Data Mining Model for Oriental Medicine VIA By Binary Independent Factor analysis", Proceedings of IEEE ,. pp.1-4.

[14]. Jenn-Lung Su, Guo-Zhen Wu, I-Pin Chao (2001). "The Approach of Data Mining Methods For Medical Database ",Proceedings of IEEE ,pp.1-3.

[15]. Safwan Mahmud Khan Md. Rafiqul Islam Morshed U. (2006). "Medical Image Classification Using an Efficient Data Mining Technique", Proceedings of IEEE, pp.1-6.

[16]. Yanwei Xing, Jie Wang and Zhihong Zhao (2007). "Combination data mining methods with new medical data to predicting outcome of Coronary Heart Disease". Pr**oceedings of** IEEE. pp.1-5.

[17]. Tsang-Hsiang Cheng, Chih-Ping Wei, Vincent S. Tseng (2009). "Feature Selection for Medical Data Mining: Comparisons of Expert Judgment and Automatic Approaches". Proceedings of IEEE. pp.1-6.

[18]. Mohammad Saraee, George Koundourakis, Babis Theodoulidis. (2007). "EasyMiner: Data Mining In Medical Databases", Proceedings of IEEE, pp.1-3.

[19]. Sam Chao(2009) "An Incremental Decision Tree Learning Methodology Regarding Attributes In Medical Data Mining". Proceedings of the Eighth International Conference on Machine Learning and Cybernetics, Baoding, pp.101-105.

[20]. My Chau Tu AND Dongil Shin (2009). "A Comparative Study of Medical Data Classification Methods Based on Decision Tree and Bagging Algorithms.", Proceedings of IEEE. pp.1-5.

[21]. Vili Podgorelec, Marjan Heriko Maribor, (2006).," Improving Mining of Medical Data by Outliers Prediction.", Proceedings of IEEE, pp.1-6.